autogluon_each_location

October 19, 2023

[1]:

```
# config

label = 'y'
metric = 'mean_absolute_error'
time_limit = None
presets = 'best_quality'

do_drop_ds = True
# hour, dayofweek, dayofmonth, month, year
use_dt_attrs = []#["hour", "year"]
use_estimated_diff_attr = False
use_is_estimated_attr = True

use_groups = False
n_groups = 8

auto_stack = False
num_stack_levels = 2
num_bag_folds = 8
num_bag_sets = 20

use_tune_data = True
use_test_data = False
tune_and_test_length = 0.5 # 3 months from end
holdout_frac = None
use_bag_holdout = True # Enable this if there is a large gap between score_val␣
↪and score_test in stack models.

sample_weight = None#'sample_weight' #None
weight_evaluation = False
sample_weight_estimated = 1
sample_weight_may_july = 1

run_analysis = False


shift_predictions_by_average_of_negatives_then_clip = False
```

```
clip_predictions = True
shift_predictions = False
```

[2]:
```
import pandas as pd
import numpy as np



import warnings
warnings.filterwarnings("ignore")


def feature_engineering(X):
    # shift all columns with "1h" in them by 1 hour, so that for index 16:00,␣
 ↪we have the values from 17:00
    # but only for the columns with "1h" in the name
    #X_shifted = X.filter(regex="\dh").shift(-1, axis=1)
    #print(f"Number of columns with 1h in name: {X_shifted.columns}")


    columns = ['clear_sky_energy_1h:J', 'diffuse_rad_1h:J', 'direct_rad_1h:J',
       'fresh_snow_12h:cm', 'fresh_snow_1h:cm', 'fresh_snow_24h:cm',
       'fresh_snow_3h:cm', 'fresh_snow_6h:cm']

    X_shifted = X[X.index.minute==0][columns].copy()
    # loop through all rows and check if index + 1 hour is in the index, if so␣
 ↪get that value, else nan
    count1 = 0
    count2 = 0
    for i in range(len(X_shifted)):
        if X_shifted.index[i] + pd.Timedelta('1 hour') in X.index:
            count1 += 1
            X_shifted.iloc[i] = X.loc[X_shifted.index[i] + pd.Timedelta('1␣
 ↪hour')][columns]
        else:
            count2 += 1
            X_shifted.iloc[i] = np.nan

    print("COUNT1", count1)
    print("COUNT2", count2)

    X_old_unshifted = X[X.index.minute==0][columns]
    # rename X_old_unshifted columns to have _not_shifted at the end
    X_old_unshifted.columns = [f"{col}_not_shifted" for col in X_old_unshifted.
 ↪columns]

    # put the shifted columns back into the original dataframe
```

```
    #X[columns] = X_shifted[columns]



    date_calc = None
    if "date_calc" in X.columns:
        date_calc = X[X.index.minute == 0]['date_calc']

    # resample to hourly
    print("index: ", X.index[0])
    X = X.resample('H').mean()
    print("index AFTER: ", X.index[0])

    X[columns] = X_shifted[columns]
    #X[X_old_unshifted.columns] = X_old_unshifted

    if date_calc is not None:
        X['date_calc'] = date_calc

    return X




def fix_X(X, name):
    # Convert 'date_forecast' to datetime format and replace original column␣
 ↪with 'ds'
    X['ds'] = pd.to_datetime(X['date_forecast'])
    X.drop(columns=['date_forecast'], inplace=True, errors='ignore')
    X.sort_values(by='ds', inplace=True)
    X.set_index('ds', inplace=True)


    X = feature_engineering(X)

    return X



def handle_features(X_train_observed, X_train_estimated, X_test, y_train):
    X_train_observed = fix_X(X_train_observed, "X_train_observed")
    X_train_estimated = fix_X(X_train_estimated, "X_train_estimated")
    X_test = fix_X(X_test, "X_test")


    if weight_evaluation:
        # add sample weights, which are 1 for observed and 3 for estimated
```

```
        X_train_observed["sample_weight"] = 1
        X_train_estimated["sample_weight"] = sample_weight_estimated
        X_test["sample_weight"] = sample_weight_estimated


    y_train['ds'] = pd.to_datetime(y_train['time'])
    y_train.drop(columns=['time'], inplace=True)
    y_train.sort_values(by='ds', inplace=True)
    y_train.set_index('ds', inplace=True)

    return X_train_observed, X_train_estimated, X_test, y_train




def preprocess_data(X_train_observed, X_train_estimated, X_test, y_train,␣
↪location):
    # convert to datetime
    X_train_observed, X_train_estimated, X_test, y_train =␣
↪handle_features(X_train_observed, X_train_estimated, X_test, y_train)

    if use_estimated_diff_attr:
        X_train_observed["estimated_diff_hours"] = 0
        X_train_estimated["estimated_diff_hours"] = (X_train_estimated.index -␣
↪pd.to_datetime(X_train_estimated["date_calc"])).dt.total_seconds() / 3600
        X_test["estimated_diff_hours"] = (X_test.index - pd.
↪to_datetime(X_test["date_calc"])).dt.total_seconds() / 3600

        X_train_estimated["estimated_diff_hours"] =␣
↪X_train_estimated["estimated_diff_hours"].astype('int64')
        # the filled once will get dropped later anyways, when we drop y nans
        X_test["estimated_diff_hours"] = X_test["estimated_diff_hours"].
↪fillna(-50).astype('int64')

    if use_is_estimated_attr:
        X_train_observed["is_estimated"] = 0
        X_train_estimated["is_estimated"] = 1
        X_test["is_estimated"] = 1

    # drop date_calc
    X_train_estimated.drop(columns=['date_calc'], inplace=True)
    X_test.drop(columns=['date_calc'], inplace=True)


    y_train["y"] = y_train["pv_measurement"].astype('float64')
    y_train.drop(columns=['pv_measurement'], inplace=True)
```

```python
    X_train = pd.concat([X_train_observed, X_train_estimated])


    # clip all y values to 0 if negative
    y_train["y"] = y_train["y"].clip(lower=0)

    X_train = pd.merge(X_train, y_train, how="inner", left_index=True,␣
 ↪right_index=True)

    # print number of nans in y
    print(f"Number of nans in y: {X_train['y'].isna().sum()}")


    X_train["location"] = location
    X_test["location"] = location

    return X_train, X_test
# Define locations
locations = ['A', 'B', 'C']

X_trains = []
X_tests = []
# Loop through locations
for loc in locations:
    print(f"Processing location {loc}...")
    # Read target training data
    y_train = pd.read_parquet(f'{loc}/train_targets.parquet')

    # Read estimated training data and add location feature
    X_train_estimated = pd.read_parquet(f'{loc}/X_train_estimated.parquet')

    # Read observed training data and add location feature
    X_train_observed= pd.read_parquet(f'{loc}/X_train_observed.parquet')

    # Read estimated test data and add location feature
    X_test_estimated = pd.read_parquet(f'{loc}/X_test_estimated.parquet')

    # Preprocess data
    X_train, X_test = preprocess_data(X_train_observed, X_train_estimated,␣
 ↪X_test_estimated, y_train, loc)

    X_trains.append(X_train)
    X_tests.append(X_test)

# Concatenate all data and save to csv
X_train = pd.concat(X_trains)
X_test = pd.concat(X_tests)
```

```
Processing location A…
COUNT1 29667
COUNT2 1
index:  2019-06-02 22:00:00
index AFTER:  2019-06-02 22:00:00
COUNT1 4392
COUNT2 2
index:  2022-10-28 22:00:00
index AFTER:  2022-10-28 22:00:00
COUNT1 702
COUNT2 18
index:  2023-05-01 00:00:00
index AFTER:  2023-05-01 00:00:00
Number of nans in y: 0
Processing location B…
COUNT1 29232
COUNT2 1
index:  2019-01-01 00:00:00
index AFTER:  2019-01-01 00:00:00
COUNT1 4392
COUNT2 2
index:  2022-10-28 22:00:00
index AFTER:  2022-10-28 22:00:00
COUNT1 702
COUNT2 18
index:  2023-05-01 00:00:00
index AFTER:  2023-05-01 00:00:00
Number of nans in y: 4
Processing location C…
COUNT1 29206
COUNT2 1
index:  2019-01-01 00:00:00
index AFTER:  2019-01-01 00:00:00
COUNT1 4392
COUNT2 2
index:  2022-10-28 22:00:00
index AFTER:  2022-10-28 22:00:00
COUNT1 702
COUNT2 18
index:  2023-05-01 00:00:00
index AFTER:  2023-05-01 00:00:00
Number of nans in y: 6059
```

1 Feature enginering

[3]:
```
import numpy as np
import pandas as pd

X_train.dropna(subset=['y', 'direct_rad_1h:J', 'diffuse_rad_1h:J'],␣
 ↪inplace=True)


for attr in use_dt_attrs:
    X_train[attr] = getattr(X_train.index, attr)
    X_test[attr] = getattr(X_test.index, attr)

#print(X_train.head())


# If the "sample_weight" column is present and weight_evaluation is True,␣
 ↪multiply sample_weight with sample_weight_may_july if the ds is between␣
 ↪05-01 00:00:00 and 07-03 23:00:00, else add sample_weight as a column to␣
 ↪X_train
if weight_evaluation:
    if "sample_weight" not in X_train.columns:
        X_train["sample_weight"] = 1

    X_train.loc[((X_train.index.month >= 5) & (X_train.index.month <= 6)) |␣
 ↪((X_train.index.month == 7) & (X_train.index.day <= 3)), "sample_weight"] *=␣
 ↪sample_weight_may_july


print(X_train.iloc[200])
print(X_train[((X_train.index.month >= 5) & (X_train.index.month <= 6)) |␣
 ↪((X_train.index.month == 7) & (X_train.index.day <= 3))].head(1))


if use_groups:
    # fix groups for cross validation
    locations = X_train['location'].unique()  # Assuming 'location' is the name␣
 ↪of the column representing locations

    grouped_dfs = []  # To store data frames split by location

    # Loop through each unique location
    for loc in locations:
        loc_df = X_train[X_train['location'] == loc]

        # Sort the DataFrame for this location by the time column
        loc_df = loc_df.sort_index()
```

```
        # Calculate the size of each group for this location
        group_size = len(loc_df) // n_groups

        # Create a new 'group' column for this location
        loc_df['group'] = np.repeat(range(n_groups),␣
 ↪repeats=[group_size]*(n_groups-1) + [len(loc_df) - group_size*(n_groups-1)])

        # Append to list of grouped DataFrames
        grouped_dfs.append(loc_df)

    # Concatenate all the grouped DataFrames back together
    X_train = pd.concat(grouped_dfs)
    X_train.sort_index(inplace=True)
    print(X_train["group"].head())




to_drop = ["snow_drift:idx", "snow_density:kgm3", "wind_speed_w_1000hPa:ms",␣
 ↪"dew_or_rime:idx", "prob_rime:p", "fresh_snow_12h:cm", "fresh_snow_24h:cm",␣
 ↪"wind_speed_u_10m:ms", "wind_speed_v_10m:ms", "snow_melt_10min:mm",␣
 ↪"rain_water:kgm2", "dew_point_2m:K", "precip_5min:mm", "absolute_humidity_2m:
 ↪gm3", "air_density_2m:kgm3"]

X_train.drop(columns=to_drop, inplace=True)
X_test.drop(columns=to_drop, inplace=True)

X_train.to_csv('X_train_raw.csv', index=True)
X_test.to_csv('X_test_raw.csv', index=True)
```

```
absolute_humidity_2m:gm3              7.825
air_density_2m:kgm3                   1.245
ceiling_height_agl:m            2085.774902
clear_sky_energy_1h:J           1685498.875
clear_sky_rad:W                  452.100006
cloud_base_agl:m                2085.774902
dew_or_rime:idx                         0.0
dew_point_2m:K                   280.549988
diffuse_rad:W                    140.800003
diffuse_rad_1h:J                 538581.625
direct_rad:W                     102.599998
direct_rad_1h:J                 439453.8125
effective_cloud_cover:p           71.849998
elevation:m                             6.0
fresh_snow_12h:cm                       0.0
```

```
fresh_snow_1h:cm                          0.0
fresh_snow_24h:cm                         0.0
fresh_snow_3h:cm                          0.0
fresh_snow_6h:cm                          0.0
is_day:idx                                1.0
is_in_shadow:idx                          0.0
msl_pressure:hPa                  1026.349976
precip_5min:mm                            0.0
precip_type_5min:idx                      0.0
pressure_100m:hPa                 1013.325012
pressure_50m:hPa                  1019.450012
prob_rime:p                               0.0
rain_water:kgm2                           0.0
relative_humidity_1000hPa:p         77.099998
sfc_pressure:hPa                  1025.550049
snow_density:kgm3                         NaN
snow_depth:cm                             0.0
snow_drift:idx                            0.0
snow_melt_10min:mm                        0.0
snow_water:kgm2                           0.0
sun_azimuth:d                       93.415253
sun_elevation:d                     27.633499
super_cooled_liquid_water:kgm2          0.025
t_1000hPa:K                           282.625
total_cloud_cover:p                 71.849998
visibility:m                        44177.875
wind_speed_10m:ms                       2.675
wind_speed_u_10m:ms                      -2.3
wind_speed_v_10m:ms                      -1.4
wind_speed_w_1000hPa:ms                   0.0
is_estimated                                0
y                                     2991.12
location                                    A
Name: 2019-06-11 06:00:00, dtype: object
                     absolute_humidity_2m:gm3  air_density_2m:kgm3  \
ds
2019-06-02 22:00:00                       7.7              1.22825

                     ceiling_height_agl:m  clear_sky_energy_1h:J  \
ds
2019-06-02 22:00:00           1728.949951                    0.0

                     clear_sky_rad:W  cloud_base_agl:m  dew_or_rime:idx  \
ds
2019-06-02 22:00:00              0.0       1728.949951              0.0

                     dew_point_2m:K  diffuse_rad:W  diffuse_rad_1h:J  …  \
ds                                                                    …
```

```
2019-06-02 22:00:00      280.299988           0.0                0.0  …

                     t_1000hPa:K  total_cloud_cover:p  visibility:m  \
ds
2019-06-02 22:00:00   286.225006                100.0  40386.476562

                     wind_speed_10m:ms  wind_speed_u_10m:ms  \
ds
2019-06-02 22:00:00                3.6               -3.575

                     wind_speed_v_10m:ms  wind_speed_w_1000hPa:ms  \
ds
2019-06-02 22:00:00                 -0.5                      0.0

                     is_estimated    y  location
ds
2019-06-02 22:00:00             0  0.0         A

[1 rows x 48 columns]
```

[5]:
```
# Create a plot of X_train showing its "y" and color it based on the value of␣
↪the sample_weight column.

#import matplotlib.pyplot as plt
#import seaborn as sns
#sns.scatterplot(data=X_train, x=X_train.index, y="y", hue="sample_weight",␣
↪palette="deep", size=3)
#plt.show()
```

[6]:
```
def normalize_sample_weights_per_location(df):
    for loc in locations:
        loc_df = df[df["location"] == loc]
        loc_df["sample_weight"] = loc_df["sample_weight"] /␣
↪loc_df["sample_weight"].sum() * loc_df.shape[0]
        df[df["location"] == loc] = loc_df
    return df

import pandas as pd
import numpy as np

def split_and_shuffle_data(input_data, num_bins, frac1):
    """
    Splits the input_data into num_bins and shuffles them, then divides the␣
↪bins into two datasets based on the given fraction for the first set.

    Args:
        input_data (pd.DataFrame): The data to be split and shuffled.
```

```
        num_bins (int): The number of bins to split the data into.
        frac1 (float): The fraction of each bin to go into the first output␣
↪dataset.

    Returns:
        pd.DataFrame, pd.DataFrame: The two output datasets.
    """
    # Validate the input fraction
    if frac1 < 0 or frac1 > 1:
        raise ValueError("frac1 must be between 0 and 1.")

    if frac1==1:
        return input_data, pd.DataFrame()

    # Calculate the fraction for the second output set
    frac2 = 1 - frac1

    # Calculate bin size
    bin_size = len(input_data) // num_bins

    # Initialize empty DataFrames for output
    output_data1 = pd.DataFrame()
    output_data2 = pd.DataFrame()

    for i in range(num_bins):
        # Shuffle the data in the current bin
        np.random.seed(i)
        current_bin = input_data.iloc[i * bin_size: (i + 1) * bin_size].
↪sample(frac=1)

        # Calculate the sizes for each output set
        size1 = int(len(current_bin) * frac1)

        # Split and append to output DataFrames
        output_data1 = pd.concat([output_data1, current_bin.iloc[:size1]])
        output_data2 = pd.concat([output_data2, current_bin.iloc[size1:]])

    # Shuffle and split the remaining data
    remaining_data = input_data.iloc[num_bins * bin_size:].sample(frac=1)
    remaining_size1 = int(len(remaining_data) * frac1)

    output_data1 = pd.concat([output_data1, remaining_data.iloc[:
↪remaining_size1]])
    output_data2 = pd.concat([output_data2, remaining_data.iloc[remaining_size1:
↪]])

    return output_data1, output_data2
```

[7]:
```
from autogluon.tabular import TabularDataset, TabularPredictor
from autogluon.timeseries import TimeSeriesDataFrame
import numpy as np
data = TabularDataset('X_train_raw.csv')
# set group column of train_data be increasing from 0 to 7 based on time, the␣
↪first 1/8 of the data is group 0, the second 1/8 of the data is group 1, etc.
data['ds'] = pd.to_datetime(data['ds'])
data = data.sort_values(by='ds')

# # print size of the group for each location
# for loc in locations:
#     print(f"Location {loc}:")
#     print(train_data[train_data["location"] == loc].groupby('group').size())


# get end date of train data and subtract 3 months
#split_time = pd.to_datetime(train_data["ds"]).max() - pd.
↪Timedelta(hours=tune_and_test_length)
# 2022-10-28 22:00:00
split_time = pd.to_datetime("2022-10-28 22:00:00")
train_set = TabularDataset(data[data["ds"] < split_time])
test_set = TabularDataset(data[data["ds"] >= split_time])

# shuffle test_set and only grab tune_and_test_length percent of it, rest goes␣
↪to train_set
test_set, new_train_set = split_and_shuffle_data(test_set, 40,␣
↪tune_and_test_length)


print("Length of train set before adding test set", len(train_set))
# add rest to train_set
train_set = pd.concat([train_set, new_train_set])
print("Length of train set after adding test set", len(train_set))
print("Length of test set", len(test_set))




if use_groups:
    test_set = test_set.drop(columns=['group'])


tuning_data = None
if use_tune_data:
    if use_test_data:
        # split test_set in half, use first half for tuning
        tuning_data, test_data = [], []
```

```
        for loc in locations:
            loc_test_set = test_set[test_set["location"] == loc]
            # randomly shuffle the loc_test_set
            loc_tuning_data, loc_test_data =␣
↪split_and_shuffle_data(loc_test_set, 40, 0.5)
            tuning_data.append(loc_tuning_data)
            test_data.append(loc_test_data)
        tuning_data = pd.concat(tuning_data)
        test_data = pd.concat(test_data)
        print("Shapes of tuning and test", tuning_data.shape[0], test_data.
↪shape[0], tuning_data.shape[0] + test_data.shape[0])

    else:
        tuning_data = test_set
        print("Shape of tuning", tuning_data.shape[0])

    # ensure sample weights for your tuning data sum to the number of rows in␣
↪the tuning data.
    if weight_evaluation:
        tuning_data = normalize_sample_weights_per_location(tuning_data)


else:
    if use_test_data:
        test_data = test_set
        print("Shape of test", test_data.shape[0])


train_data = train_set

# ensure sample weights for your training (or tuning) data sum to the number of␣
↪rows in the training (or tuning) data.
if weight_evaluation:
    train_data = normalize_sample_weights_per_location(train_data)
    if use_test_data:
        test_data = normalize_sample_weights_per_location(test_data)


train_data = TabularDataset(train_data)
if use_tune_data:
    tuning_data = TabularDataset(tuning_data)
if use_test_data:
    test_data = TabularDataset(test_data)
```

```
Length of train set before adding test set 82026
Length of train set after adding test set 87486
Length of test set 5459
```

```
Shape of tuning 5459
```

[8]:
```
if run_analysis:
    import autogluon.eda.auto as auto
    auto.dataset_overview(train_data=train_data, test_data=test_data,␣
↪label="y", sample=None)
```

[9]:
```
if run_analysis:
    auto.target_analysis(train_data=train_data, label="y", sample=None)
```

2 Starting

[10]:
```
import os


# Get the last submission number
last_submission_number = int(max([int(filename.split('_')[1].split('.')[0]) for␣
↪filename in os.listdir('submissions') if "submission" in filename]))
print("Last submission number:", last_submission_number)
print("Now creating submission number:", last_submission_number + 1)

# Create the new filename
new_filename = f'submission_{last_submission_number + 1}'

hello = os.environ.get('HELLO')
if hello is not None:
    new_filename += f'_{hello}'

print("New filename:", new_filename)
```

```
Last submission number: 97
Now creating submission number: 98
New filename: submission_98
```

[11]:
```
predictors = [None, None, None]
```

[12]:
```
def fit_predictor_for_location(loc):
    print(f"Training model for location {loc}...")
    # sum of sample weights for this location, and number of rows, for both␣
↪train and tune data and test data
    if weight_evaluation:
        print("Train data sample weight sum:",␣
↪train_data[train_data["location"] == loc]["sample_weight"].sum())
        print("Train data number of rows:", train_data[train_data["location"]␣
↪== loc].shape[0])
        if use_tune_data:
```

```
            print("Tune data sample weight sum:",␣
↪tuning_data[tuning_data["location"] == loc]["sample_weight"].sum())
            print("Tune data number of rows:",␣
↪tuning_data[tuning_data["location"] == loc].shape[0])
        if use_test_data:
            print("Test data sample weight sum:",␣
↪test_data[test_data["location"] == loc]["sample_weight"].sum())
            print("Test data number of rows:", test_data[test_data["location"]␣
↪== loc].shape[0])
    predictor = TabularPredictor(
        label=label,
        eval_metric=metric,
        path=f"AutogluonModels/{new_filename}_{loc}",
        # sample_weight=sample_weight,
        # weight_evaluation=weight_evaluation,
        # groups="group" if use_groups else None,
    ).fit(
        train_data=train_data[train_data["location"] == loc].
↪drop(columns=["ds"]),
        time_limit=time_limit,
        # presets=presets,
        num_stack_levels=num_stack_levels,
        num_bag_folds=num_bag_folds if not use_groups else 2,# just put␣
↪somethin, will be overwritten anyways
        num_bag_sets=num_bag_sets,
        tuning_data=tuning_data[tuning_data["location"] == loc].
↪reset_index(drop=True).drop(columns=["ds"]) if use_tune_data else None,
        use_bag_holdout=use_bag_holdout,
        # holdout_frac=holdout_frac,
    )

    # evaluate on test data
    if use_test_data:
        # drop sample_weight column
        t = test_data[test_data["location"] == loc]#.
↪drop(columns=["sample_weight"])
        perf = predictor.evaluate(t)
        print("Evaluation on test data:")
        print(perf[predictor.eval_metric.name])

    return predictor

loc = "A"
predictors[0] = fit_predictor_for_location(loc)
```

```
Warning: path already exists! This predictor may overwrite an existing
predictor! path="AutogluonModels/submission_98_A"
```

```
Beginning AutoGluon training …
AutoGluon will save models to "AutogluonModels/submission_98_A/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   186.17 GB / 315.93 GB (58.9%)
Train Data Rows:    31872
Train Data Columns: 32
Tuning Data Rows:    2187
Tuning Data Columns: 32
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and many unique label-values observed).
        Label info (max, min, mean, stddev): (5733.42, 0.0, 649.68162,
1178.37671)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    131876.83 MB
        Train Data (Original)  Memory Usage: 10.42 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 1 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…

Training model for location A…

        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…
        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
        Useless Original Features (Count: 2): ['elevation:m', 'location']
                These features carry no predictive signal and should be manually
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
```

```
                ('float', []) : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', [])   :  1 | ['is_estimated']
        Types of features in processed data (raw dtype, special dtypes):
                ('float', [])     : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', ['bool']) :  1 | ['is_estimated']
        0.1s = Fit runtime
        30 features in original data used to generate 30 features in processed
data.
        Train Data (Processed) Memory Usage: 7.94 MB (0.0% of available memory)
Data preprocessing and feature engineering runtime = 0.16s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
use_bag_holdout=True, will use tuning_data as holdout (will not be used for
early stopping).
User-specified model hyperparameters to be fit:
{
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
AutoGluon will fit 3 stack levels (L1 to L3) …
Fitting 11 L1 models …
Fitting model: KNeighborsUnif_BAG_L1 …
        -132.836         = Validation score   (-mean_absolute_error)
        0.03s    = Training   runtime
        0.42s    = Validation runtime
```

```
Fitting model: KNeighborsDist_BAG_L1 …
        -132.5631        = Validation score   (-mean_absolute_error)
        0.02s    = Training   runtime
        0.39s    = Validation runtime
Fitting model: LightGBMXT_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -93.8449         = Validation score   (-mean_absolute_error)
        430.79s  = Training   runtime
        2777.24s         = Validation runtime
Fitting model: LightGBM_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -97.2345         = Validation score   (-mean_absolute_error)
        282.99s  = Training   runtime
        836.39s  = Validation runtime
Fitting model: RandomForestMSE_BAG_L1 …
        -105.3243        = Validation score   (-mean_absolute_error)
        7.8s     = Training   runtime
        1.13s    = Validation runtime
Fitting model: CatBoost_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -102.9612        = Validation score   (-mean_absolute_error)
        2160.56s         = Training   runtime
        11.06s   = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L1 …
        -107.644         = Validation score   (-mean_absolute_error)
        1.55s    = Training   runtime
        1.13s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -107.0868        = Validation score   (-mean_absolute_error)
        299.36s  = Training   runtime
        35.37s   = Validation runtime
Fitting model: XGBoost_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -103.034         = Validation score   (-mean_absolute_error)
        196.35s  = Training   runtime
        50.73s   = Validation runtime
Fitting model: NeuralNetTorch_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -94.6002         = Validation score   (-mean_absolute_error)
        696.28s  = Training   runtime
        10.64s   = Validation runtime
```

```
Fitting model: LightGBMLarge_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -95.9861         = Validation score   (-mean_absolute_error)
        1081.35s         = Training   runtime
        3246.57s         = Validation runtime
Fitting model: WeightedEnsemble_L2 …
        -90.1813         = Validation score   (-mean_absolute_error)
        0.6s     = Training   runtime
        0.0s     = Validation runtime
Fitting 9 L2 models …
Fitting model: LightGBMXT_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.5317         = Validation score   (-mean_absolute_error)
        50.08s   = Training   runtime
        20.58s   = Validation runtime
Fitting model: LightGBM_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -91.5851         = Validation score   (-mean_absolute_error)
        39.29s   = Training   runtime
        11.26s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L2 …
        -93.4758         = Validation score   (-mean_absolute_error)
        13.47s   = Training   runtime
        1.29s    = Validation runtime
Fitting model: CatBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.9194         = Validation score   (-mean_absolute_error)
        97.5s    = Training   runtime
        6.92s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L2 …
        -93.5771         = Validation score   (-mean_absolute_error)
        2.28s    = Training   runtime
        1.27s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -91.9844         = Validation score   (-mean_absolute_error)
        304.79s  = Training   runtime
        31.23s   = Validation runtime
Fitting model: XGBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.6837         = Validation score   (-mean_absolute_error)
        47.91s   = Training   runtime
```

```
        8.8s     = Validation runtime
Fitting model: NeuralNetTorch_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -88.3813         = Validation score   (-mean_absolute_error)
        459.26s  = Training   runtime
        15.73s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -91.3817         = Validation score   (-mean_absolute_error)
        83.05s   = Training   runtime
        22.53s   = Validation runtime
Fitting model: WeightedEnsemble_L3 …
        -87.9869         = Validation score   (-mean_absolute_error)
        0.36s    = Training   runtime
        0.0s     = Validation runtime
Fitting 9 L3 models …
Fitting model: LightGBMXT_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.0198         = Validation score   (-mean_absolute_error)
        37.09s   = Training   runtime
        14.47s   = Validation runtime
Fitting model: LightGBM_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -91.1946         = Validation score   (-mean_absolute_error)
        39.55s   = Training   runtime
        11.96s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L3 …
        -93.4164         = Validation score   (-mean_absolute_error)
        12.5s    = Training   runtime
        1.25s    = Validation runtime
Fitting model: CatBoost_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.6612         = Validation score   (-mean_absolute_error)
        77.79s   = Training   runtime
        6.73s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L3 …
        -93.9341         = Validation score   (-mean_absolute_error)
        2.04s    = Training   runtime
        1.25s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -93.4845         = Validation score   (-mean_absolute_error)
```

```
        305.09s  = Training   runtime
        28.8s    = Validation runtime
Fitting model: XGBoost_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -91.3641         = Validation score   (-mean_absolute_error)
        45.76s   = Training   runtime
        8.0s     = Validation runtime
Fitting model: NeuralNetTorch_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -87.886  = Validation score   (-mean_absolute_error)
        302.65s  = Training   runtime
        15.14s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -90.3168         = Validation score   (-mean_absolute_error)
        79.43s   = Training   runtime
        20.98s   = Validation runtime
Fitting model: WeightedEnsemble_L4 …
        -87.5955         = Validation score   (-mean_absolute_error)
        0.35s    = Training   runtime
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 7516.39s … Best model:
"WeightedEnsemble_L4"
TabularPredictor saved. To load, use: predictor =
TabularPredictor.load("AutogluonModels/submission_98_A/")
```

[13]:

```
import matplotlib.pyplot as plt

leaderboards = [None, None, None]
def leaderboard_for_location(i, loc):
    if use_test_data:
        lb = predictors[i].leaderboard(test_data[test_data["location"] == loc])
        lb["location"] = loc
        plt.scatter(test_data[test_data["location"] == loc]["y"].index,␣
 ↪test_data[test_data["location"] == loc]["y"])
        if use_tune_data:
            plt.scatter(tuning_data[tuning_data["location"] == loc]["y"].index,␣
 ↪tuning_data[tuning_data["location"] == loc]["y"])
        plt.show()

        return lb
    else:
        return pd.DataFrame()
```

```
leaderboards[0] = leaderboard_for_location(0, loc)
```

[14]:
```
loc = "B"
predictors[1] = fit_predictor_for_location(loc)
leaderboards[1] = leaderboard_for_location(1, loc)
```

```
Beginning AutoGluon training …
AutoGluon will save models to "AutogluonModels/submission_98_B/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   155.84 GB / 315.93 GB (49.3%)
Train Data Rows:    31020
Train Data Columns: 32
Tuning Data Rows:    1797
Tuning Data Columns: 32
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and many unique label-values observed).
        Label info (max, min, mean, stddev): (1152.3, -0.0, 99.56591, 196.469)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    128461.83 MB
        Train Data (Original)  Memory Usage: 10.04 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 1 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…
        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…
        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
        Useless Original Features (Count: 2): ['elevation:m', 'location']
```

```
Training model for location B…
```

```
                These features carry no predictive signal and should be manually
```

```
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
                ('float', []) : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', [])   :  1 | ['is_estimated']
        Types of features in processed data (raw dtype, special dtypes):
                ('float', [])     : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', ['bool']) :  1 | ['is_estimated']
        0.1s = Fit runtime
        30 features in original data used to generate 30 features in processed
data.
        Train Data (Processed) Memory Usage: 7.65 MB (0.0% of available memory)
Data preprocessing and feature engineering runtime = 0.18s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
use_bag_holdout=True, will use tuning_data as holdout (will not be used for
early stopping).
User-specified model hyperparameters to be fit:
{
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
AutoGluon will fit 3 stack levels (L1 to L3) …
```

```
Fitting 11 L1 models …
Fitting model: KNeighborsUnif_BAG_L1 …
        -26.8714         = Validation score   (-mean_absolute_error)
        0.02s    = Training   runtime
        0.63s    = Validation runtime
Fitting model: KNeighborsDist_BAG_L1 …
        -26.8453         = Validation score   (-mean_absolute_error)
        0.02s    = Training   runtime
        0.41s    = Validation runtime
Fitting model: LightGBMXT_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -16.9704         = Validation score   (-mean_absolute_error)
        428.03s  = Training   runtime
        2774.3s  = Validation runtime
Fitting model: LightGBM_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -16.4977         = Validation score   (-mean_absolute_error)
        467.12s  = Training   runtime
        2634.62s         = Validation runtime
Fitting model: RandomForestMSE_BAG_L1 …
        -17.8185         = Validation score   (-mean_absolute_error)
        8.59s    = Training   runtime
        1.12s    = Validation runtime
Fitting model: CatBoost_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -17.3306         = Validation score   (-mean_absolute_error)
        2261.6s  = Training   runtime
        11.7s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L1 …
        -17.1018         = Validation score   (-mean_absolute_error)
        1.48s    = Training   runtime
        1.13s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.9093         = Validation score   (-mean_absolute_error)
        294.87s  = Training   runtime
        30.3s    = Validation runtime
Fitting model: XGBoost_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -16.5663         = Validation score   (-mean_absolute_error)
        1226.19s         = Training   runtime
        3341.58s         = Validation runtime
Fitting model: NeuralNetTorch_BAG_L1 …
```

```
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -12.7362         = Validation score   (-mean_absolute_error)
        1206.1s  = Training   runtime
        10.4s    = Validation runtime
Fitting model: LightGBMLarge_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -15.42   = Validation score   (-mean_absolute_error)
        1106.42s         = Training   runtime
        3520.94s         = Validation runtime
Fitting model: WeightedEnsemble_L2 …
        -12.7362         = Validation score   (-mean_absolute_error)
        0.41s    = Training   runtime
        0.0s     = Validation runtime
Fitting 9 L2 models …
Fitting model: LightGBMXT_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.7073         = Validation score   (-mean_absolute_error)
        84.03s   = Training   runtime
        72.97s   = Validation runtime
Fitting model: LightGBM_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.0619         = Validation score   (-mean_absolute_error)
        47.32s   = Training   runtime
        16.49s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L2 …
        -14.2502         = Validation score   (-mean_absolute_error)
        13.81s   = Training   runtime
        1.32s    = Validation runtime
Fitting model: CatBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.2332         = Validation score   (-mean_absolute_error)
        250.74s  = Training   runtime
        6.27s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L2 …
        -13.8692         = Validation score   (-mean_absolute_error)
        2.0s     = Training   runtime
        1.16s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.0843         = Validation score   (-mean_absolute_error)
        311.99s  = Training   runtime
        28.95s   = Validation runtime
```

```
Fitting model: XGBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.0994         = Validation score   (-mean_absolute_error)
        55.45s   = Training   runtime
        11.49s   = Validation runtime
Fitting model: NeuralNetTorch_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -12.9121         = Validation score   (-mean_absolute_error)
        768.03s  = Training   runtime
        15.62s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.8287         = Validation score   (-mean_absolute_error)
        194.24s  = Training   runtime
        56.19s   = Validation runtime
Fitting model: WeightedEnsemble_L3 …
        -12.9051         = Validation score   (-mean_absolute_error)
        0.35s    = Training   runtime
        0.0s     = Validation runtime
Fitting 9 L3 models …
Fitting model: LightGBMXT_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.3859         = Validation score   (-mean_absolute_error)
        43.73s   = Training   runtime
        18.82s   = Validation runtime
Fitting model: LightGBM_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.0854         = Validation score   (-mean_absolute_error)
        40.33s   = Training   runtime
        11.74s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L3 …
        -14.1727         = Validation score   (-mean_absolute_error)
        13.07s   = Training   runtime
        1.19s    = Validation runtime
Fitting model: CatBoost_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.8511         = Validation score   (-mean_absolute_error)
        102.4s   = Training   runtime
        6.01s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L3 …
        -14.1536         = Validation score   (-mean_absolute_error)
        1.92s    = Training   runtime
```

```
        1.19s    = Validation runtime
Fitting model: NeuralNetFastAI_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.2388         = Validation score   (-mean_absolute_error)
        309.76s  = Training   runtime
        29.43s   = Validation runtime
Fitting model: XGBoost_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.9943         = Validation score   (-mean_absolute_error)
        47.21s   = Training   runtime
        8.91s    = Validation runtime
Fitting model: NeuralNetTorch_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.0114         = Validation score   (-mean_absolute_error)
        349.63s  = Training   runtime
        14.83s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L3 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.9044         = Validation score   (-mean_absolute_error)
        91.33s   = Training   runtime
        27.09s   = Validation runtime
Fitting model: WeightedEnsemble_L4 …
        -13.0114         = Validation score   (-mean_absolute_error)
        0.35s    = Training   runtime
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 10181.03s … Best model:
"WeightedEnsemble_L2"
TabularPredictor saved. To load, use: predictor =
TabularPredictor.load("AutogluonModels/submission_98_B/")
```

[ ]:
```
loc = "C"
predictors[2] = fit_predictor_for_location(loc)
leaderboards[2] = leaderboard_for_location(2, loc)
```

```
Beginning AutoGluon training …
AutoGluon will save models to "AutogluonModels/submission_98_C/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   113.16 GB / 315.93 GB (35.8%)
Train Data Rows:    24594
Train Data Columns: 32
```

```
Tuning Data Rows:    1475
Tuning Data Columns: 32
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and label-values can't be converted to int).
        Label info (max, min, mean, stddev): (999.6, -0.0, 79.8926, 168.407)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    126801.54 MB
        Train Data (Original)  Memory Usage: 7.98 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 1 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…
        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…
        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
        Useless Original Features (Count: 2): ['elevation:m', 'location']
                These features carry no predictive signal and should be manually
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
                ('float', []) : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', [])   :  1 | ['is_estimated']
        Types of features in processed data (raw dtype, special dtypes):
                ('float', [])     : 29 | ['ceiling_height_agl:m',
'clear_sky_energy_1h:J', 'clear_sky_rad:W', 'cloud_base_agl:m', 'diffuse_rad:W',
…]
                ('int', ['bool']) :  1 | ['is_estimated']
        0.1s = Fit runtime
        30 features in original data used to generate 30 features in processed
data.
        Train Data (Processed) Memory Usage: 6.07 MB (0.0% of available memory)
```

```
Data preprocessing and feature engineering runtime = 0.15s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
use_bag_holdout=True, will use tuning_data as holdout (will not be used for
early stopping).
User-specified model hyperparameters to be fit:
{
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
AutoGluon will fit 3 stack levels (L1 to L3) …
Fitting 11 L1 models …
Fitting model: KNeighborsUnif_BAG_L1 …

Training model for location C…

        -23.875  = Validation score   (-mean_absolute_error)
        0.02s    = Training   runtime
        0.27s    = Validation runtime
Fitting model: KNeighborsDist_BAG_L1 …
        -23.8095         = Validation score   (-mean_absolute_error)
        0.02s    = Training   runtime
        0.31s    = Validation runtime
Fitting model: LightGBMXT_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -14.3864         = Validation score   (-mean_absolute_error)
        250.82s  = Training   runtime
        30.18s   = Validation runtime
Fitting model: XGBoost_BAG_L1 …
```

```
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.5867         = Validation score   (-mean_absolute_error)
        525.13s  = Training   runtime
        888.91s  = Validation runtime
Fitting model: NeuralNetTorch_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -13.602  = Validation score   (-mean_absolute_error)
        549.41s  = Training   runtime
        10.08s   = Validation runtime
Fitting model: LightGBMLarge_BAG_L1 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -12.6629         = Validation score   (-mean_absolute_error)
        989.74s  = Training   runtime
        2331.3s  = Validation runtime
Fitting model: WeightedEnsemble_L2 …
        -11.7585         = Validation score   (-mean_absolute_error)
        0.43s    = Training   runtime
        0.0s     = Validation runtime
Fitting 9 L2 models …
Fitting model: LightGBMXT_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -11.939  = Validation score   (-mean_absolute_error)
        45.49s   = Training   runtime
        16.39s   = Validation runtime
Fitting model: LightGBM_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -11.8245         = Validation score   (-mean_absolute_error)
        37.26s   = Training   runtime
        11.35s   = Validation runtime
Fitting model: RandomForestMSE_BAG_L2 …
        -11.9945         = Validation score   (-mean_absolute_error)
        8.25s    = Training   runtime
        0.77s    = Validation runtime
Fitting model: CatBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -12.3639         = Validation score   (-mean_absolute_error)
        105.53s  = Training   runtime
        4.63s    = Validation runtime
Fitting model: ExtraTreesMSE_BAG_L2 …
        -11.7564         = Validation score   (-mean_absolute_error)
        1.24s    = Training   runtime
        0.78s    = Validation runtime
```

```
Fitting model: NeuralNetFastAI_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -11.8692         = Validation score   (-mean_absolute_error)
        253.61s  = Training   runtime
        28.72s   = Validation runtime
Fitting model: XGBoost_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
        -11.7169         = Validation score   (-mean_absolute_error)
        48.43s   = Training   runtime
        8.49s    = Validation runtime
Fitting model: NeuralNetTorch_BAG_L2 …
        Fitting 160 child models (S1F1 - S20F8) | Fitting with
ParallelLocalFoldFittingStrategy
```

[ ]:
```
# save leaderboards to csv
pd.concat(leaderboards).to_csv(f"leaderboards/{new_filename}.csv")
```

3 Submit

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt

future_test_data = TabularDataset('X_test_raw.csv')
future_test_data["ds"] = pd.to_datetime(future_test_data["ds"])
#test_data
```

[ ]:
```
test_ids = TabularDataset('test.csv')
test_ids["time"] = pd.to_datetime(test_ids["time"])
# merge test_data with test_ids
future_test_data_merged = pd.merge(future_test_data, test_ids, how="inner",␣
↪right_on=["time", "location"], left_on=["ds", "location"])

#test_data_merged
```

[ ]:
```
# predict, grouped by location
predictions = []
location_map = {
    "A": 0,
    "B": 1,
    "C": 2
}
for loc, group in future_test_data.groupby('location'):
    i = location_map[loc]
    subset = future_test_data_merged[future_test_data_merged["location"] ==␣
↪loc].reset_index(drop=True)
```

```
    #print(subset)
    pred = predictors[i].predict(subset)
    subset["prediction"] = pred
    predictions.append(subset)

    # get past predictions
    train_data.loc[train_data["location"] == loc, "prediction"] = ␣
↪predictors[i].predict(train_data[train_data["location"] == loc])
    if use_tune_data:
        tuning_data.loc[tuning_data["location"] == loc, "prediction"] = ␣
↪predictors[i].predict(tuning_data[tuning_data["location"] == loc])
    if use_test_data:
        test_data.loc[test_data["location"] == loc, "prediction"] = ␣
↪predictors[i].predict(test_data[test_data["location"] == loc])
```

[ ]:

```
# plot predictions for location A, in addition to train data for A
for loc, idx in location_map.items():
    fig, ax = plt.subplots(figsize=(20, 10))
    # plot train data
    train_data[train_data["location"]==loc].plot(x='ds', y='y', ax=ax,␣
↪label="train data")
    if use_tune_data:
        tuning_data[tuning_data["location"]==loc].plot(x='ds', y='y', ax=ax,␣
↪label="tune data")
    if use_test_data:
        test_data[test_data["location"]==loc].plot(x='ds', y='y', ax=ax,␣
↪label="test data")

    # plot predictions
    predictions[idx].plot(x='ds', y='prediction', ax=ax, label="predictions")

    # plot past predictions
    #train_data_with_dates[train_data_with_dates["location"]==loc].plot(x='ds',␣
↪y='prediction', ax=ax, label="past predictions")
    train_data[train_data["location"]==loc].plot(x='ds', y='prediction', ax=ax,␣
↪label="past predictions train")
    if use_tune_data:
        tuning_data[tuning_data["location"]==loc].plot(x='ds', y='prediction',␣
↪ax=ax, label="past predictions tune")
    if use_test_data:
        test_data[test_data["location"]==loc].plot(x='ds', y='prediction',␣
↪ax=ax, label="past predictions test")


    # title
    ax.set_title(f"Predictions for location {loc}")
```

[ ]:
```python
temp_predictions = [prediction.copy() for prediction in predictions]
if clip_predictions:
    # clip predictions smaller than 0 to 0
    for pred in temp_predictions:
        # print smallest prediction
        print("Smallest prediction:", pred["prediction"].min())
        pred.loc[pred["prediction"] < 0, "prediction"] = 0
        print("Smallest prediction after clipping:", pred["prediction"].min())

# Instead of clipping, shift all prediction values up by the largest negative␣
↪number.
# This way, the smallest prediction will be 0.
elif shift_predictions:
    for pred in temp_predictions:
        # print smallest prediction
        print("Smallest prediction:", pred["prediction"].min())
        pred["prediction"] = pred["prediction"] - pred["prediction"].min()
        print("Smallest prediction after clipping:", pred["prediction"].min())

elif shift_predictions_by_average_of_negatives_then_clip:
    for pred in temp_predictions:
        # print smallest prediction
        print("Smallest prediction:", pred["prediction"].min())
        mean_negative = pred[pred["prediction"] < 0]["prediction"].mean()
        # if not nan
        if mean_negative == mean_negative:
            pred["prediction"] = pred["prediction"] - mean_negative

        pred.loc[pred["prediction"] < 0, "prediction"] = 0
        print("Smallest prediction after clipping:", pred["prediction"].min())



# concatenate predictions
submissions_df = pd.concat(temp_predictions)
submissions_df = submissions_df[["id", "prediction"]]
submissions_df
```

[ ]:
```python
# Save the submission DataFrame to submissions folder, create new name based on␣
↪last submission, format is submission_<last_submission_number + 1>.csv

# Save the submission
print(f"Saving submission to submissions/{new_filename}.csv")
submissions_df.to_csv(os.path.join('submissions', f"{new_filename}.csv"),␣
↪index=False)
print("jall1a")
```

[ ]:
```python
train_data_with_dates = TabularDataset('X_train_raw.csv')
train_data_with_dates["ds"] = pd.to_datetime(train_data_with_dates["ds"])
# feature importance
location="A"
split_time = pd.Timestamp("2022-10-28 22:00:00")
estimated = train_data_with_dates[train_data_with_dates["ds"] >= split_time]
estimated = estimated[estimated["location"] == location]
predictors[0].feature_importance(feature_stage="original", data=estimated,␣
↪time_limit=60*10)
```

[ ]:
```python
# feature importance
observed = train_data_with_dates[train_data_with_dates["ds"] < split_time]
observed = observed[observed["location"] == location]
predictors[0].feature_importance(feature_stage="original", data=observed,␣
↪time_limit=60*10)
```

[ ]:
```python
# save this running notebook
from IPython.display import display, Javascript
import time

# hei123

display(Javascript("IPython.notebook.save_checkpoint();"))

time.sleep(3)
```

[ ]:
```python
# save this notebook to submissions folder
import subprocess
import os
subprocess.run(["jupyter", "nbconvert", "--to", "pdf", "--output", os.path.
↪join('notebook_pdfs', f"{new_filename}.pdf"), "autogluon_each_location.
↪ipynb"])
```

[ ]:
```python
# display(Javascript("IPython.notebook.save_checkpoint();"))
# time.sleep(3)

# subprocess.run(["jupyter", "nbconvert", "--to", "pdf", "--output", os.path.
↪join('notebook_pdfs', f"{new_filename}_with_feature_importance.pdf"),␣
↪"autogluon_each_location.ipynb"])
```

[ ]:
```python
# import subprocess

# def execute_git_command(directory, command):
#     """Execute a Git command in the specified directory."""
#     try:
#         result = subprocess.check_output(['git', '-C', directory] + command,␣
↪stderr=subprocess.STDOUT)
```

```
#         return result.decode('utf-8').strip(), True
#     except subprocess.CalledProcessError as e:
#         print(f"Git command failed with message: {e.output.decode('utf-8').
↪strip()}")
#         return e.output.decode('utf-8').strip(), False

# git_repo_path = "."

# execute_git_command(git_repo_path, ['config', 'user.email',␣
↪'henrikskog01@gmail.com'])
# execute_git_command(git_repo_path, ['config', 'user.name', hello if hello is␣
↪not None else 'Henrik eller Jørgen'])

# branch_name = new_filename

# # add datetime to branch name
# branch_name += f"_{pd.Timestamp.now().strftime('%Y-%m-%d_%H-%M-%S')}"

# commit_msg = "run result"

# execute_git_command(git_repo_path, ['checkout', '-b',branch_name])

# # Navigate to your repo and commit changes
# execute_git_command(git_repo_path, ['add', '.'])
# execute_git_command(git_repo_path, ['commit', '-m',commit_msg])

# # Push to remote
# output, success = execute_git_command(git_repo_path, ['push',␣
↪'origin',branch_name])

# # If the push fails, try setting an upstream branch and push again
# if not success and 'upstream' in output:
#     print("Attempting to set upstream and push again...")
#     execute_git_command(git_repo_path, ['push', '--set-upstream',␣
↪'origin',branch_name])
#     execute_git_command(git_repo_path, ['push', 'origin', 'henrik_branch'])

# execute_git_command(git_repo_path, ['checkout', 'main'])
```